
Towards a Useful and Usable Bioinformatics Infrastructure

Terence Critchlow

*Center for Applied Scientific Computing
Lawrence Livermore National Laboratory
www.llnl.gov/CASC/people/critchlow*

**Georgia Tech
Department of Computer Science**

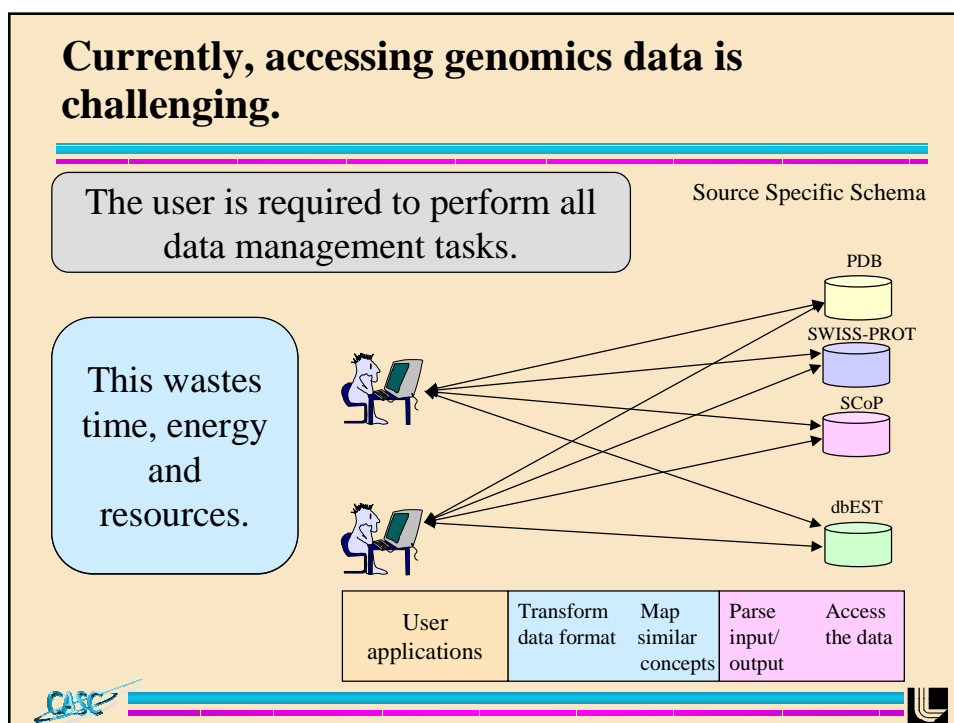


Outline

- **A day in the life of a geneticist/biologist.**
- **CS to the rescue! (well, maybe....)**
- **DataFoundry's meta-data infrastructure.**
- **Future research directions.**

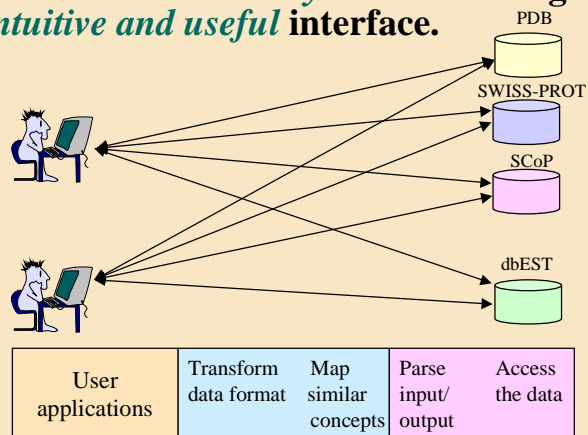


Man, this sure would be easier if I could move between these different sites without having to reformat the data every time.



What is our ideal environment?

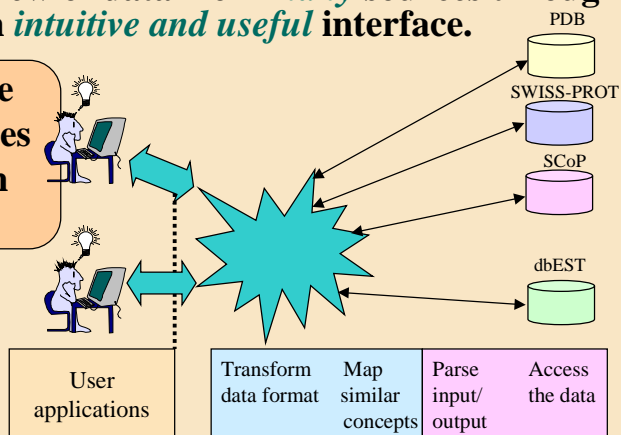
A *single* location that provides *effective* access to a *consistent* view of data from *many* sources through an *intuitive and useful* interface.



What is our ideal environment?

A *single* location that provides *effective* access to a *consistent* view of data from *many* sources through an *intuitive and useful* interface.

Businesses use data warehouses to accomplish this.

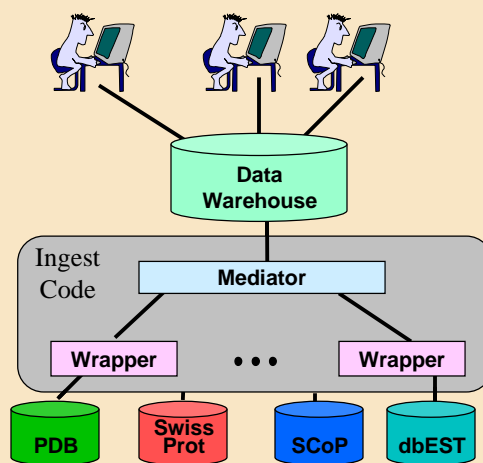


**CS can address these issues
(or can it?)**

CASC



Data warehouses



- **Interfaces**

- ✿ provide intuitive access to the data
- ✿ possibly change data format to meet user expectations

- **Warehouse**

- ✿ stores a consistent view of data in a local repository

- **Mediator**

- ✿ transform data from source format to warehouse format

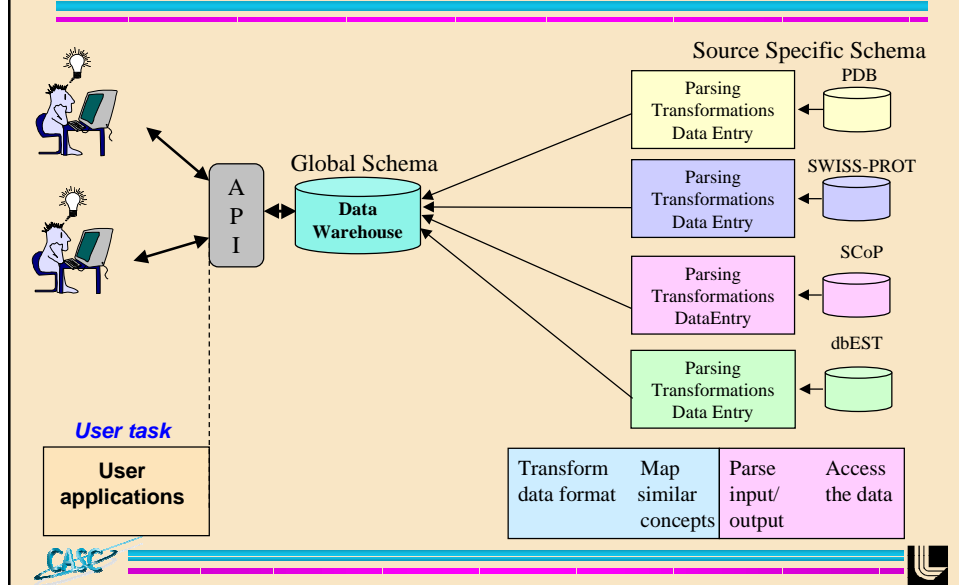
- **Wrappers**

- ✿ read data from source into internal representation

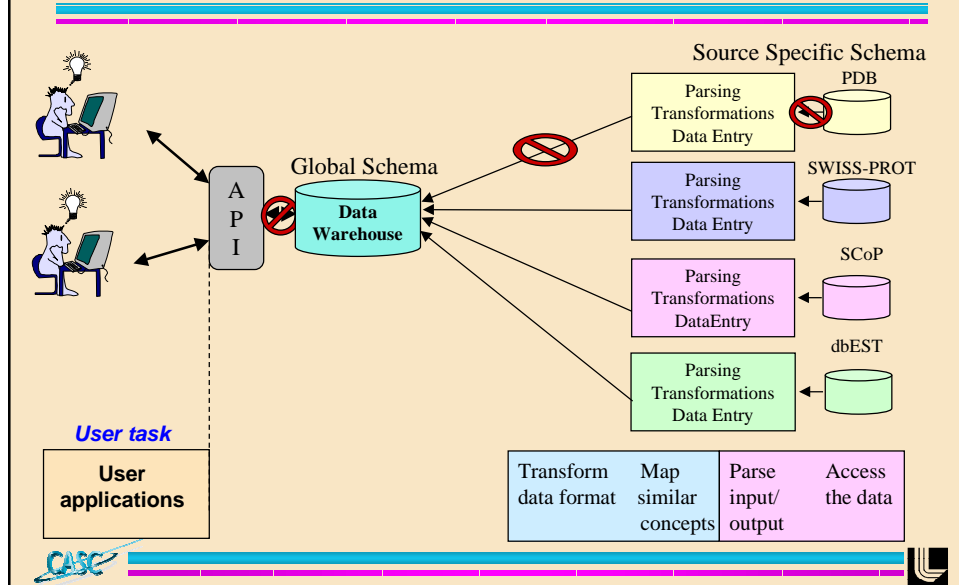
CASC



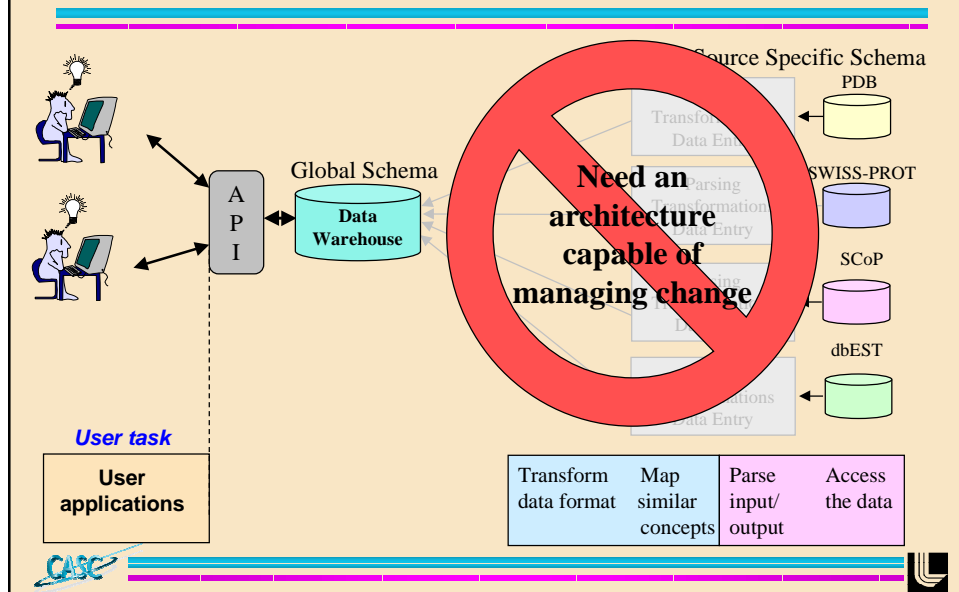
Typical approaches combine wrapper and mediator functionality.



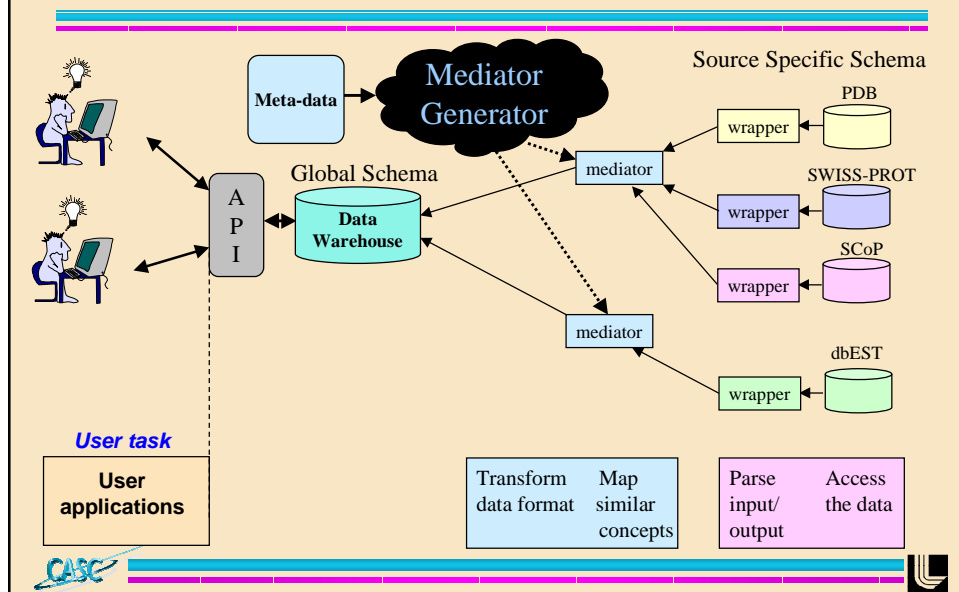
When a source changes, propagating the change can be time consuming.



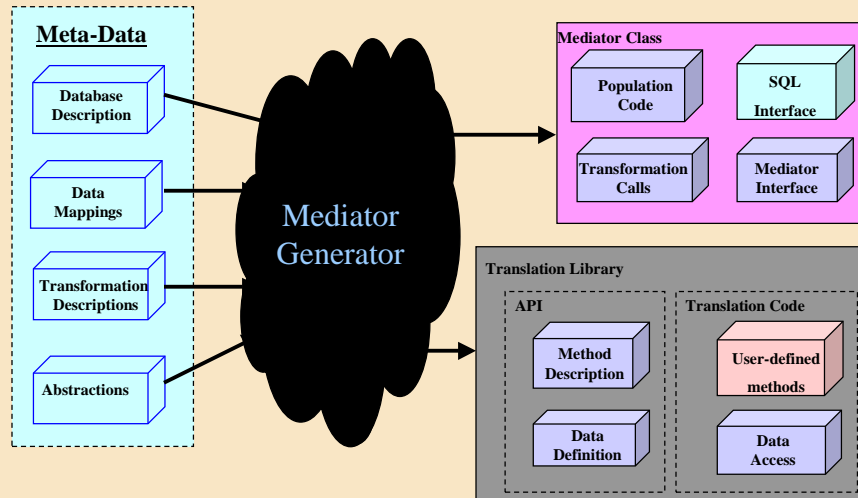
When a source changes, propagating the change can be time consuming.



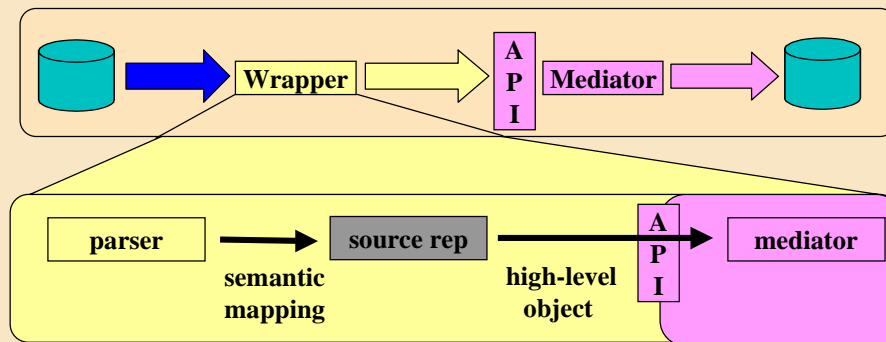
DataFoundry separates these tasks.



We use meta-data to automatically generate mediators.



The translation library and mediator class are used by the wrapper



Behind the scenes, the effect of using the DataFoundry technology is dramatic.

Cost of integrating SCoP into warehouse that already contains PDB and SWISS-PROT.

Activity/ integration style	typical	DataFoundry	diff	%diff
understanding SCOP	2.0	2.0	0.0	0
writing wrapper	4.5	2.5	2.0	44%
modifying schema	0.5	0.5	0.0	0
writing mediator	4.0	0.0	4.0	---
modifying meta-data	0.0	1.0	(1.0)	---
total time in days	11.0	6.0	5.0	45%

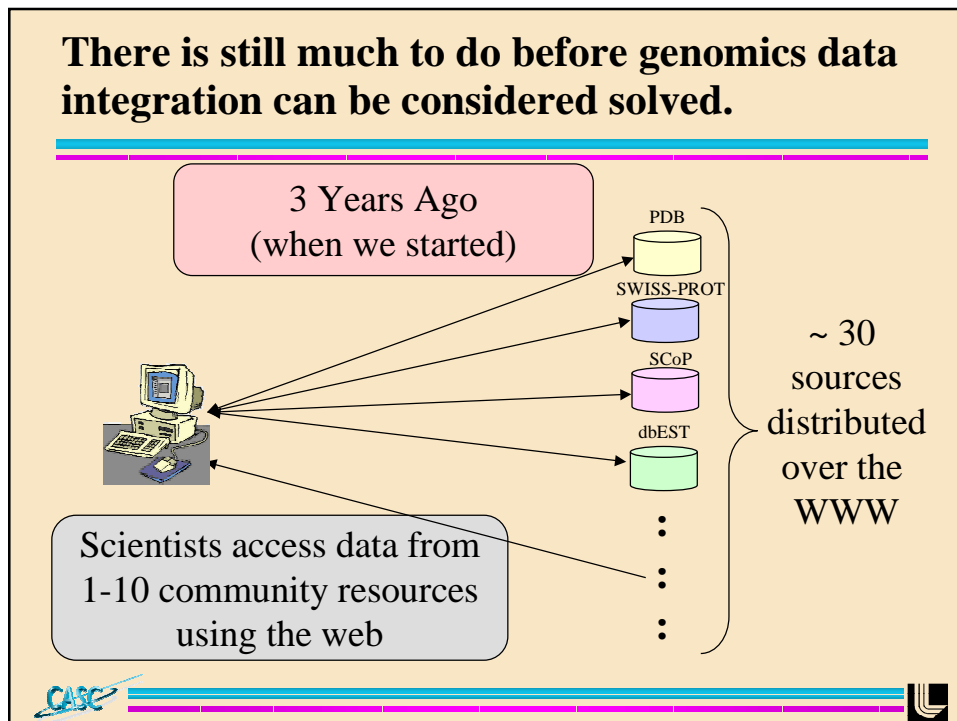
DataFoundry makes it possible to build and maintain scientific data warehouses.



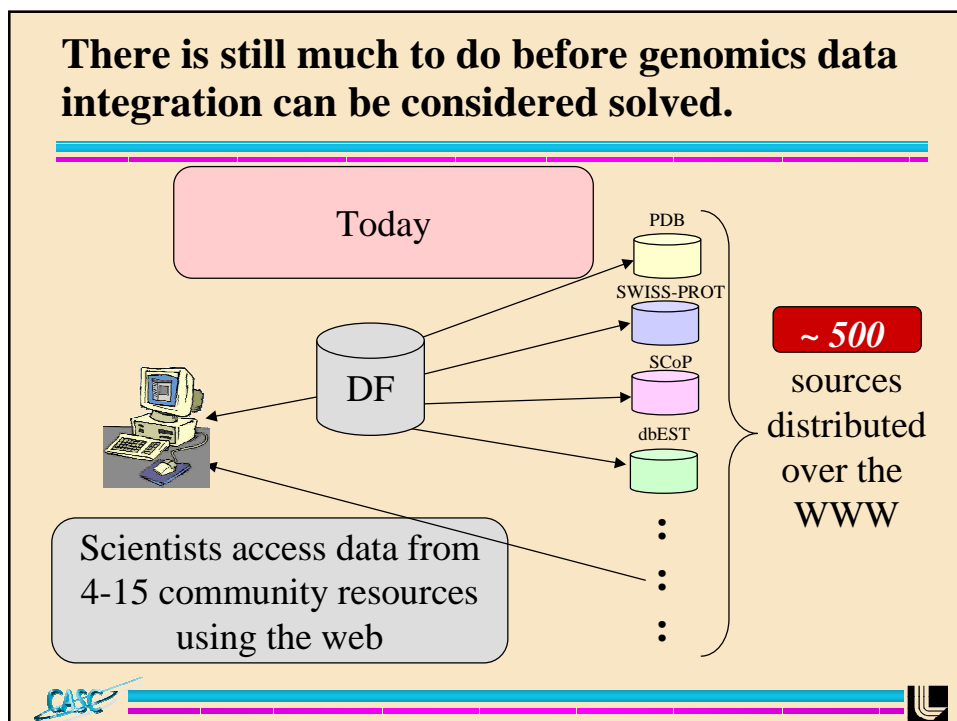
**We are continue to explore
data management issues on
two fronts.**



There is still much to do before genomics data integration can be considered solved.



There is still much to do before genomics data integration can be considered solved.

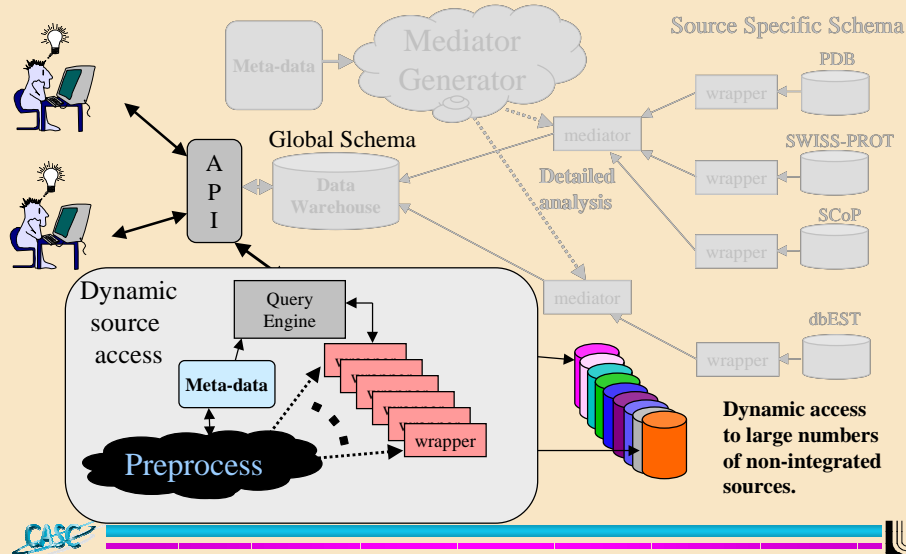


There are several challenges to querying hundreds of sources:

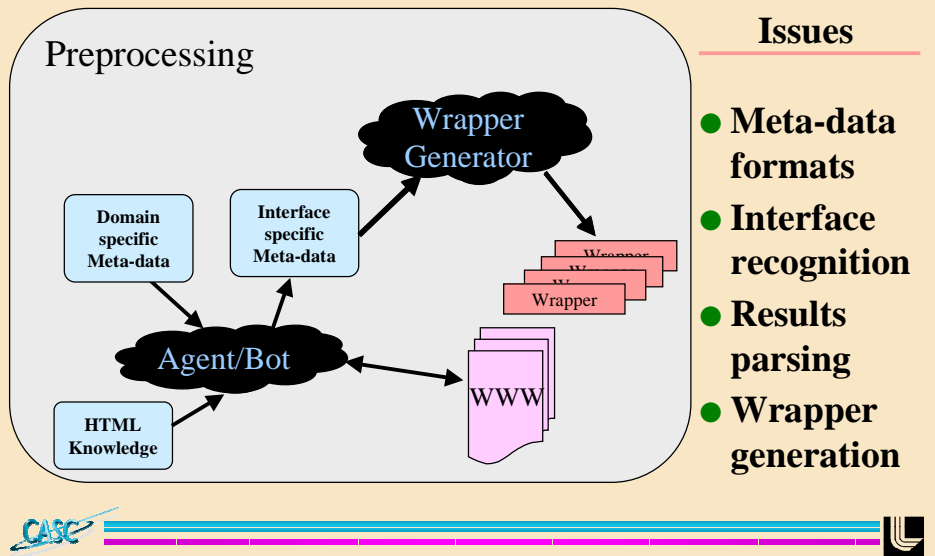
- Usable descriptions of relevant interfaces
 - ✿ query input/output descriptions required
- Automatic creation of wrappers
 - ✿ human interaction infeasible
- Intuitive presentation of non-integrated data
 - ✿ semantic integration of data too difficult
- Selective querying of data sources
 - ✿ data source categorization needed



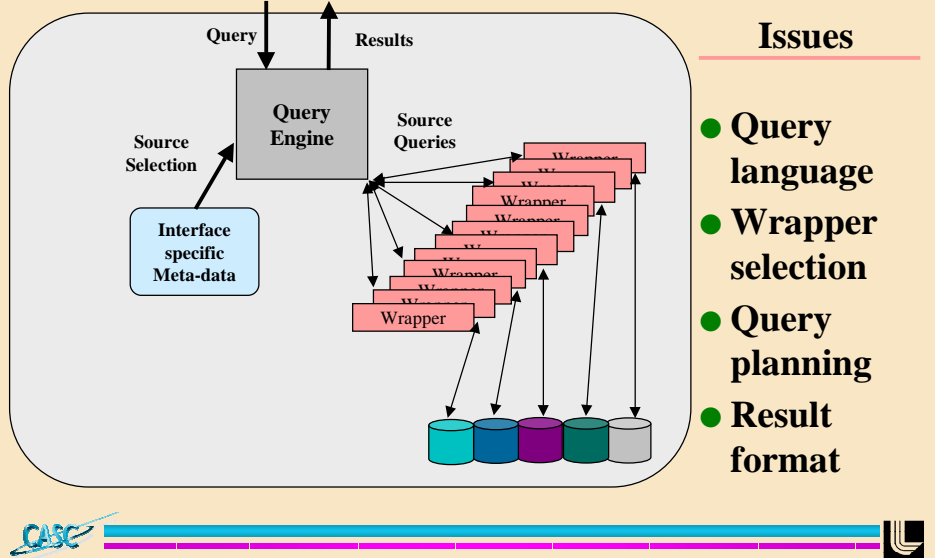
A hybrid approach to information access holds the most promise.



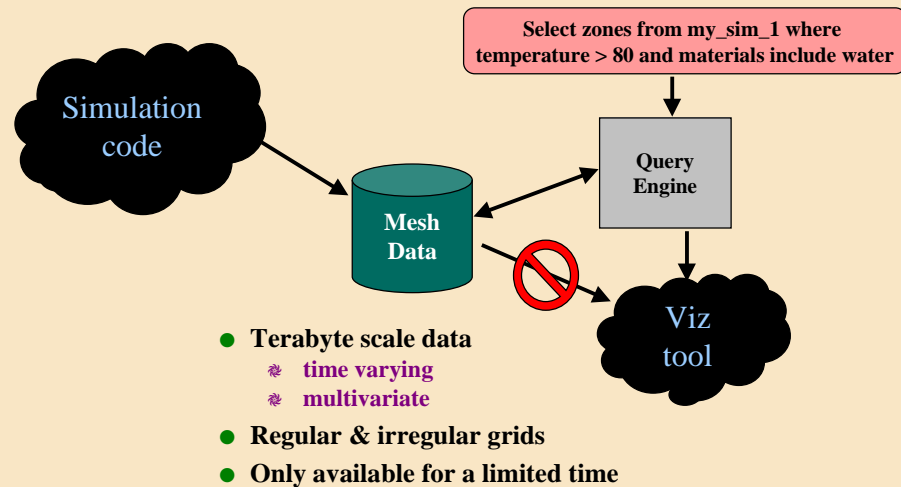
A preprocessing step is used to automatically generate wrappers.



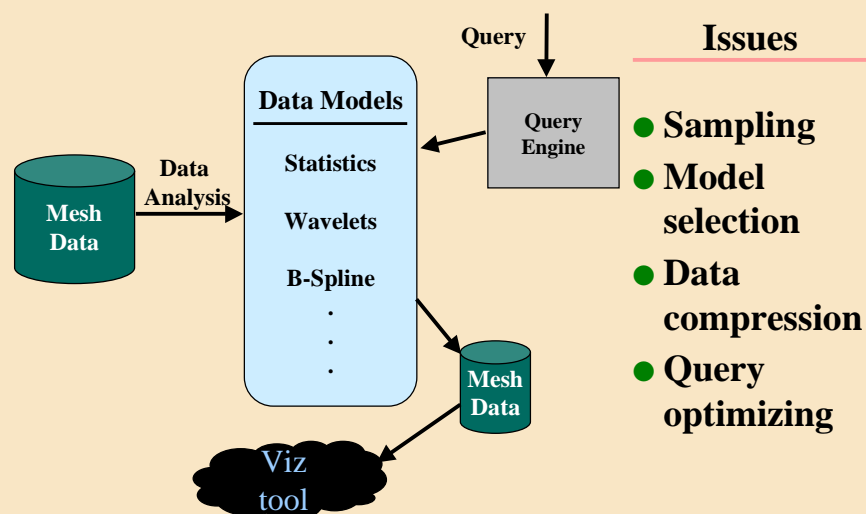
The query engine uses the resulting meta-data and wrappers.



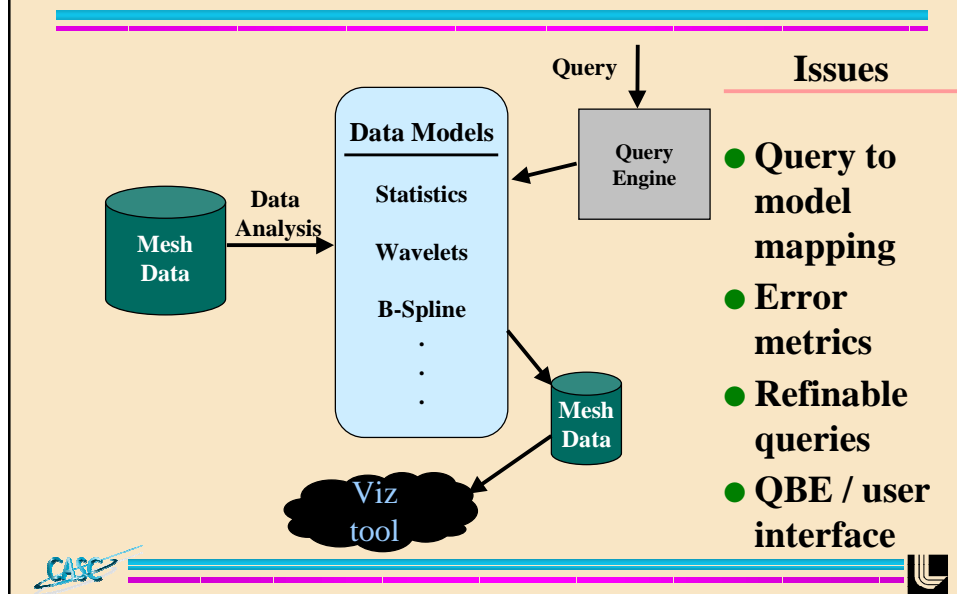
Our second research thrust is querying large-scale simulation data.



Again, preprocessing is an important step.



Again, preprocessing is an important step.



Current status

Both projects are very new.
Work has only recently begun and is ongoing.

Data Integration

- Strawman design of meta-data formats completed

✿ focus of upcoming workshop

Ad-hoc Queries


- Prototype infrastructure with canned queries nearing completion

✿ currently ignoring most research issues



Questions?

www.llnl.gov/CASC/people/critchlow



This work was performed under the auspices of the U.S.
Department of Energy by University of California Lawrence
Livermore National Laboratory under contract No. W-7405-
ENG-48.

